图书情报工作
第62卷 第21期 2018年11月

融入术语知识的专利主题发现方法*

■ 俞琰[1,2] 赵乃瑄[1]

[1] 南京工业大学信息服务部 南京 210009 [2] 东南大学成贤学院电子与计算机学院 南京 211816

摘要：[目的/意义]针对专利主题分析中以词为基本单位会造成专利中的多词术语难以被识别、主题模型结果不佳的问题，提出融入术语的专利主题发现模型，以解决该问题。[方法/过程]模型首先引入类别熵，有效地识别出专利文献中的术语；然后利用泛化波利亚瓮模型增加语义相似术语分配到同一主题的概率，以缓解术语作为基本主题模型分析单位所带来的数据稀疏性问题。[结果/结论]实验结果表明本文提出的模型包含的术语信息提高了主题生成的质量，使主题表示具有更强的可读性和主题判别性。

关键词：专利分析 主题发现 术语

分类号：G202

DOI：10.13266/j.issn.0252-3116.2018.21.015

1 引言

专利文献中蕴含着丰富的各个领域问题的解决方案，有效的专利文献分析能够判断领域技术热点、识别领域核心技术、预测领域技术发展趋势、帮助研发人员从中获得启发与借鉴，从而缩短创新设计时间、节约创新设计经费。因此，专利文献分析具有重要的研究意义。不同于传统的文献分析方法，主题模型通过分析文献集合中词语共现的概率分布，能够挖掘文献中隐藏的语义信息。随着主题模型的广泛应用，研究者尝试将主题模型应用于专利文献分析之中，以揭示专利文献深层次知识结构[1-10]。大量研究结果表明基于主题模型的专利文献分析具有较强的实践意义。

然而，专利主题模型分析的基本单位是词，造成专利中的多词术语难以被识别，导致主题表意不清、不易理解等现象。例如，在中文专利主题分析之前，需要首先进行分词处理。依靠一个相对完整的分词词典，目前的中文分词技术基本能够满足研究需要。然而，专利文献与通用语料相比，包含大量没有被分词词典记录的词组型术语，称之为未登录词，这些未登录词造成中文专利文献分词效果不尽理想，产生大量的语义碎片。例如，术语"渗硼剂"被分词工具切分为"渗 硼 剂"；术语"热浸镀"被切分为"热 浸 镀"；术语"双相不锈钢"被切分为"双相 不锈钢"。术语集中体现了领域的核心知识，而这些被切分成碎片的术语往往难以被识别而无法揭示其核心知识。此外，术语被切分成多个单词后会引起额外的共现，使得生成的主题可能出现一些无关词汇，导致主题模型结果不佳。改善主题语义的一个方法是关注比词汇更高阶的语义单元。一般是在主题模型中将传统的词分布替换为高阶语义单元的分布。然而，目前基于术语的专利主题分析还未被深入研究，其他相关领域的研究存在模型过于复杂、不易扩展、术语识别准确率低、数据共现稀疏性等问题。

针对专利主题模型分析中的上述问题，本文提出融入术语知识的专利主题发现模型。该模型将术语作为主题模型分析的基本单位，首先根据专利文献的特点，引入类别熵，有效地识别出专利文献中的术语；然后利用泛化波利亚瓮模型增加语义相似术语分配到同一主题的概率，以缓解术语作为基本主题模型分析单位所带来的数据稀疏性问题。实验结果表明，同传统的专利主题模型相比，融入术语知识的专利主题模型包含更加丰富的语义信息、具有更强的可读性和主题判别性。

2 相关研究

不同于传统的专利文本分析方法，专利主题模型

* 本文系教育部人文社会科学规划项目"大数据时代技能知识图谱构建研究"（项目编号：16YJAZH073）和国家社会科学基金一般规划项目"大数据时代支持创新设计的多维度多层次专利文本挖掘研究"（项目编号：17BTQ059）研究成果之一。

作者简介：俞琰（ORCID：0000-0002-9654-8614），副教授，博士，E-mail：yuyanyuyan2004@126.com；赵乃瑄（ORCID：0000-0001-9072-7315），馆长，教授，博士。

收稿日期：2018-04-07 修回日期：2018-06-20 本文起止页码：118-126 本文责任编辑：易飞

通过分析专利文本集合中词语共现的概率分布,挖掘专利文本隐含的语义信息,以揭示专利文本深层次知识结构。例如,J. Tang 等[1]提出一个主题驱动的专利分析方法,以分析专利竞争对手发展状况等。B. Wang 等[2]从专利标题和摘要中抽取技术术语,基于技术术语,通过添加机构信息,使用扩展的主题模型,以分析某具体领域的研究热点和方向、竞争对手企业的技术位置。H. Chen 等[3]使用主题模型评估专利权力要求书中隐藏的主题。M. Kim 等[4]使用主题模型分析专利摘要和专利权力要求书,生成专利开发地图,以理解技术的发展趋势。A. Suominen 等[5]使用主题模型分类专利文本数据,以预测将来的技术趋势。范宇等[6]利用主题模型将专利文本在词汇空间的高维表达转换到在主题空间的低维表达,实现对专利文本描述的降维,进而采用 K 近邻算法对专利文本进行聚类。王博等[7]将主题模型应用于专利文本分析领域,实现专利主题划分,解决以往专利主题分类过于粗泛、时效性差、缺乏科学性等问题。吴菲菲等[8]提取专利摘要中的名词和名词短语,构建主题模型,揭示专利文本中隐含的主题演化,最终实现企业战略动态变化挖掘。廖列法和勒孚刚[9]提出基于主题模型和分类号的专利技术演化模型。陈亮等[10]采用主题模型,从专利语料库中抽取层次主题模型,描述隐藏在专利文本中的技术结构,进行专利技术演化分析。

虽然专利主题模型取得了不错的分析效果,但是专利主题模型分析以词为单位。而专利中由多词构成的术语集中体现了领域的核心知识,这些被切分成碎片的术语往往难以被识别,并引起额外的共现,使得生成的主题可能出现一些无关词汇,导致主题模型结果不佳。改善主题语义的一个方法是关注比词更高阶的语义单元,一般是在主题模型中将传统的词分布替换为高阶语义单元的分布。根据术语识别在主题发现中实施的不同阶段,现有相关方法可大致分为术语预先处理模型、联合模型、术语后处理模型等三大类。术语预先处理模型首先进行文本术语识别任务,然后基于术语识别主题。联合模型是指模型同时识别术语和发现主题[11-13]。术语后处理模型首先按传统方式,通过处理一元主题模型得到主题模型,然后将一元词组合成术语[14]。其中,虽然实验表明联合模型能够提高主题模型的准确度,推断出的主题词含义更丰富,但是,联合模型通常构造复杂,经常面临计算复杂度高和大文档集难以扩展等问题;术语后处理模型虽然将术语识别任务与主题发现任务分开进行,降低了计算复杂度和难以扩展的问题,但是这类方法使用的一元主题模型不能保证术语中的词具有相同的主题,而这在主题术语挖掘研究中恰恰是重要部分。

因此,本文主要集中于术语预先处理模型的研究。术语预先处理模型首先进行文本术语识别任务,然后基于术语识别主题。术语识别通常使用频次或结合词性规则的方法。基于频次的方法通常进行频繁项集挖掘和关联规则挖掘[10, 15-16],将频次小于人工设定阈值的词定义为低频词,并将这些低频词去除,抽取频繁项,并统计其出现次数。基于词性规则的方法根据专利文献的构词特征,提出一套适用于专利文本的名词抽取规则,抽取名词短语构成专利主题词,然后根据这些词串出现的频次形成术语[7]。在经过对文本集合的预处理后,文档被分割成术语集合,在主题模型中,术语中的单词共享相同的潜在主题模型。

术语预先处理模型方法将术语识别任务与主题发现任务分开进行,降低了计算复杂度和难以扩展的问题,并解决了术语中单词主题不一致的问题。但是,该类方法仍存在一些不足:①在术语识别阶段,基于频次产生的术语质量不高,高频词组不一定是术语,非高频词组也可能是术语。②在主题发现阶段,未关注如何解决高阶语义单元共现稀疏的问题。主题模型核心思想一般基于文档内元素的共现[17],根据自然语言中分布的幂律特点[18],大部分词共现稀疏。进行主题分析时,相较于将文档看做词袋,把文档作为术语的集合,文档所包含元素将变得更为稀疏,元素间共现度也会进一步下降,这会对主题模型带来较大的副作用。这种稀疏性必然给主题模型学习带来一定难度。

本文提出的方法属于第三类术语预先处理模型,首先是识别术语,然后发现主题,具有简单直观、易于扩展的特点。根据中文专利的特点,引入类别熵以识别出中文专利中的术语,并引入基于泛化波利亚瓮模型,利用相关术语采样来缓解稀疏性。

3 融入术语知识的专利主题模型

本文针对专利文献的特点,提出融入术语知识的专利主题模型。模型将获取的专利文本进行分词、数据清洗等预处理之后,进行术语识别(第 3.1 部分具体介绍),然后基于识别的术语,进行基于术语的主题建模(第 3.2 部分具体介绍),以实现主题发现。

3.1 术语识别

仅仅依赖于词组出现频次,不能辨别是否是专利术语,如,专利中“发明 涉及”出现的频次远高于术语。

与普通文献相比,专利文献具有其自身特点。通常,专利文献包含通用词和术语两类。通用词通常具有主题无关性,在多个类别中均匀出现,常引出术语以作为术语与下文的衔接;而术语则表达某个领域知识,具有较高的领域相关性,在某一类别中高频出现,而在其他类别中低频出现,甚至不出现。如图 1 所示,来自 3 个不同类别的专利文本语句,经过切分工具切分后,“渗 硼剂”“电解 金属锰”和“多级 离心泵”分别是不同类别的术语;而“本”“发明”“涉及”“一种”“的”“方法”等则是通用词,在 3 个类别中均有出现。相较于一般语料,通用词和术语在专利文献中的界限更加清晰。且通用词更容易被识别。因此,本文将首先识别通用词选取候选术语,然后对候选术语进行排序,评估其成为术语的可能性。

类别 1: 本 发明 涉及 一种 渗 硼 剂 的 制备 方法。
类别 2: 本 发明 涉及 一种 电解 金属锰 的 生产 方法。
类别 3: 本 发明 涉及 一种 多级 离心泵 的 设计 方法。

图 1 专利文献通用词与术语示例

3.1.1 候选术语选取 专利中的通用词通常具有主题无关性,在各类别间以及某一类别内通常均匀出现;而术语则在某一类别或几个类别中不均匀出现。为此,本文引入包含多个类别的辅助专利文本集,使用类别间熵和类别内熵衡量词的分布情况。信息熵是信息论中重要的概念,用来度量信息的不确定程度。

具体地,设 c_1, c_2,……,c_m 为 m 个类别的辅助专利文本集,每个类别包含若干个相关专利文本,将词 w 在不同类别间的分布称为类别间信息熵(entropy categories, EC),计算公式如式(1)所示:

$$EC(w)=\sum_{j=1}^{m}\frac{df(w,c_j)}{df(w)}\times\log\frac{df(w,c_j)}{df(w)} \quad 式(1)$$

其中,EC(w)表示词 w 的类别间信息熵;df(w,c_j)表示词 w 在类别 c_j 中的文档频次;$df(w)=\sum_{j=1}^{n}df(w,c_j)$,表示词 w 在辅助专利文本集中的文档频次。由定义可知,当词只出现在单个类别的文本中时,类别间信息熵最小;当词在所有类别中均匀分布时,类别间信息熵达到最大值。由式(1)可见,类别熵 EC 越大,表明词在各类别间分布越均匀,越可能是专利通用词。

按照类别熵对词降序排列,选取大于阈值的词作为通用词,对分词后的目标专利文本进行粗切分,本文遵循[1,5,19-20],将频次大于 1 次且长度大于 1 的词串及其子串作为候选术语。

3.1.2 候选术语排序 C-value 统计量可用于计算每个候选术语成为术语的可能性[21]。它是针对术语词频计算的一种改进,可以增进嵌套多词术语的抽取,排除一些非术语词汇的干扰。当子串和母串短语同时纳入候选领域术语集时,即存在嵌套串,则可以计算其在语料中的 C-value 值来判断其是否为真正的术语。C-value 方法简单、适用性强、领域无关,考虑了候选术语的嵌套性和长度,在术语识别方面表现较好[22-27]。

具体地,C-value 值计算利用候选词的以下 4 个统计特征:候选串在语料中出现的总频次、候选串作为嵌套串在语料中出现的频次、包含该嵌套串的母串的个数、候选串的长度。其计算公式如式(2)所示:

$$C\text{-}vahue(x)=\begin{cases}\log|x|\times tf(x) & x\text{ 未被嵌套}\\ \log|x|\times\left(tf(x)-\frac{1}{p(T_x)}\sum_{y\in T_x}ft(y)\right) & \text{其他}\end{cases} \quad 式(2)$$

其中,x 表示候选术语;$|x|$表示 x 的长度;$tf(x)$表示 x 在目标专利文本集中出现的频次;T_x 表示目标专利文本集包含 x 的候选术语;p(T_x)表示目标专利文本集包含 T_x 中元素个数。由公式可知,C-value 与该候选术语在语料中的频次有关,频次越高,其术语度越大。在此基础上,又考虑了候选术语的长度,认为长串出现频次比短串出现频次更有意义,是术语的可能性更大。

3.2 主题发现

本文基于传统的主题发现模型,引入泛化波利亚瓮(Generalized Pólya Urn,GPU)模型[28],利用相关术语采样来缓解稀疏性。在传统主题发现模型基础上,引入 GPU 模型,以缓解数据稀疏性问题。

3.2.1 传统主题模型 LDA(Latent Dirichlet Allocation)模型[29]是一种常用的主题模型,由于其参数简单,不产生过拟合现象,逐渐成为主题模型的研究热点。因此本文使用 LDA 模型对专利文本进行主题建模。LDA 是一个三层贝叶斯概率模型,由词、主题和文本三层构成。该模型假设每个文本包含若干隐含主题,每个主题包含特定的词。文本和词间的关系通过隐含主题体现。隐含主题之间是相互独立的,这些主题被文本集中所有文本所共享,而每个文本有一个特定的主题分布。模型通常采用 Gibbs 采样推理方法估计主题的后验分布,计算如公式(3)[29]所示:

俞琰，赵乃瑄.融入术语知识的专利主题发现方法[J].图书情报工作,2018,62(21):118－126.

$$p(z_n^{(d)}=k|z_{-d,n},W,\alpha,\beta)\propto\frac{N_{w_{(h)}^d|k}+\beta}{N_k+V\beta}\times\frac{C_{k|d}+\alpha}{C_d+K\alpha}$$

式（3）

其中，$z_n^{(d)}=k$ 表示文本 d 中第 n 个词 w 指定为变量 k；$z_{-d,n}$表示排除文本 d 中的第 n 个词 w 指定的主题；W 表示文本集合中所有的词条；K 表示主题数，V 表示集合中总的词语数。一旦获得每个文本中每个词的主题，就可以得到 LDA 模型中 θ 和 φ 的后验估计值，计算如公式(4)[30]和公式(5)[30]所示：

$$\theta_{d,k}=\frac{N_{k|d}+\alpha}{N_d+K\alpha}$$ 式(4)

$$\varphi_{k,w}=\frac{N_{w_{(n)}^d|k}+\beta}{N_k+V\beta}$$ 式(5)

其中，$\theta_{d,k}$表示文本 d 包含主题 k 的概率；$\varphi_{k,w}$表示主题 k 中词语 w 的概率。

3.2.2　基于 GPU 的主题模型　泛化波利亚瓮(Generalized Pólya Urn，GPU)模型是一种概率统计模型，是标准波利亚翁的扩展版(Pólya Urn，PU)。在主题模型中，感兴趣的物体在翁中被表示为彩色小球，在翁中观察到一个小球属于不同颜色的概率值可以通过采样和放回操作，用瓮中各个颜色的占比近似模拟。在 PU 模型中，一个小球从瓮中随机抽出，观察其颜色；然后将其放回到瓮中，并将相同颜色的附加球添加到罐子中，并重复选择过程。这个过程类似于在 LDA 模型中使用 Gibbs 采样方法对模型进行求导。在 GPU 模型中，同样有采样和放回操作。当从瓮中随机采样出一个小球后，将其放回到瓮中，并将一个与之相同颜色的小球放回瓮中。除此之外，还要准备一些颜色相似的小球，同样放回到瓮中。因此，GPU 模型可以保证后续采样中不仅能以更高的概率再次取到相同颜色的小球，同时也能以较高的概率取到其他颜色相似的小球。这个操作可以类比到基于术语的主题模型之中，利用 GPU 模型提高术语和术语(如“奥氏体 不锈钢”和“不锈钢”)被分配相同主题的概率，从而缓解基本模型中术语共现所存在的稀疏性问题。近期的一些研究工作表明利用 GPU 模型融入先验知识是一种直接有效的方法[31－35]。

具体地，基于术语的中文专利主题发现模型中，基于 GPU 模型的 Gibbs 采样推理方法估计主题的后验分布，计算如公式(6)所示：

$$p(z_n^{(d)}=k|z_{-d,n}W,\alpha,\beta,A)\propto\frac{\sum_v N_{v|k}A_{v,t_n^{(d)}}+\beta}{\sum_{v'}\sum_v N_{v|k}A_{v,v'}+V'\beta}\times\frac{N_{k|d}+\alpha}{N_d+K\alpha}$$ 式(6)

其中，$z_n^{(d)}=k$ 表示文本 d 中第 n 个术语 t 指定为变量 k；$z_{-d,n}$表示排除文本 d 中的第 n 个术语 t 指定的主题；V'表示集合中总的术语数；A 表示大小为$|V'|\times|V'|$的 GPU 促进量矩阵，为了增大语义相近元素分配到相同主题的概率，可以通过术语的相似度对矩阵 A 进行如下赋值。矩阵 A 中的元素 $A_{v,v'}=\mathrm{sim}(v,v')$，表示术语 v 和术语 v'间的相似度，定义如式(7)所示：

$$A_{v,v'}=\begin{cases}1, & v=v'\\ sim(v,v')=\dfrac{|v\cap v'|}{|v\cup v'|}, & sim(v,v')>\sigma \text{ 并且 } v'\neq v\\ 0, & \text{其他}\end{cases}$$

式(7)

其中，$|v\cap v'|$表示术语 v 与 v'相同词个数，$|v\cup v'|$表示术语 v 和 v'不同词个数，σ 为指定的相似度阈值。阈值的设置可以过滤掉相似度低的术语对，认为其相似性不明显，不加入到 GPU 促进量矩阵 A 中，例如，术语“奥氏体 不锈钢”和“抗菌 不锈钢”的相似度 1/3。本文选取相似度阈值为 0.2。

相应地，基于 GPU 主题模型中 θ 和 φ 的后验估计值，计算如公式(8)和(9)所示：

$$\theta_{d,k}=\frac{N_{k|d}+\alpha}{N_d+K\alpha}$$ 式(8)

$$\varphi_{k,w}=\frac{\sum_v N_{v|k}A_{v,t_n^{(d)}}+\beta}{\sum_{v'}\sum_v N_{v|k}A_{v,v'}+V'\beta}$$ 式（9）

在此模型中，通过引入术语相似度知识，利用 GPU 模型提高术语和术语被分配相同主题的概率，从而缓解基本模型中术语共现所存在的稀疏性问题。当 A 为单位矩阵时，即还原为 Pólya Urn 模型。由相似的对称性可知，矩阵 A 具有稀疏性。因此，GPU 模型直接作用于采样过程并不会增加模型的复杂度和推理的难度。

4　实验

4.1　数据集

为了验证本文提出模型的有效性，本文分别选取稀土钢和电解锰两个领域的专利文献进行实验。稀土钢一般指在钢中添加一定成分的稀土元素，从而提高钢的横向性能、耐磨性能和耐腐蚀性优异性能的钢种。随着近几年钢铁行业去产能和转型升级的推动，发挥稀土资源优势，提高稀土钢的市场地位成为重要课题。金属锰在钢铁工业中的用量仅次于铁，是一种重要合金元素，在钢铁冶炼，特别是锰代镍型不锈钢和高级特殊钢的冶炼中起着非常重要的作用。此外，金属锰还在有色金属、磁性材料、催化剂、电池材料等领域扮演

重要的角色。通过对我国电解锰领域专利的系统分析,探讨我国电解锰工业的研究现状与技术发展趋势,实现合理规划与科学决策。

实验基于中国国家知识产权局专利数据库,检索范围为中国发明公开专利,检索日期为 2017 年 12 月 8 日,分别以“稀土钢”和“电解锰”为关键词进行检索,分别获得 1 547 和 1 544 条中国发明公开专利。通过数据抓取、清洗、去重后,最终分别将专利标题和摘要作为待分析的领域专利文本集。

为了获取专利通用词,根据专利 IPC 分类号,分别从 A-H 中随机抽取 2 000 个中国发明公开专利文献的标题和摘要作为辅助专利文本集。数据集基本信息如表 1 所示:

表 1 数据集基本信息

数据集类型	领域	去重后文本数(条)
目标专利文本集	稀土钢	1 511
	电解锰	1 577
辅助专利文本集	A 人类生活必需(农、轻、医)	2 000
	B 作业;运输	2 000
	C 化学;冶金	2 000
	D 纺织;造纸	2 000
	E 固定建筑物(建筑、采矿)	2 000
	F 机械工程;照明;加热;武器;爆破	2 000
	G 物理	2 000
	H 电学	2 000

4.2 评估方法

实验主要对术语识别的准确率和主题模型的质量进行评估。

首先采用 P@N 方法评估专利术语识别的准确性,即判断最终排序候选术语表中前 N 条术语的准确率。被模型自动抽取的前 N 条术语采用人工方式进行判断。为了避免主观性和领域知识的局限性,对于明显正确或错误的被识别术语直接标记相应标记,而对于很难辨别正确性的被识别术语则利用百度百科、维基、互动百科等知识网站查找是否存在对应的词条,以判别被识别术语的正确性。计算公式如式(10)所示:

$$P@N=\frac{\#\text{前 }N\text{ 条候选术语中正确的术语}}{\#\text{前 }N\text{ 条候选术语}}\times 100\% \qquad \text{式(10)}$$

接着,实验借助主题与主题的平均 KL(Kullback-Leribler)距离评估生成主题的质量。KL 距离常用来衡量两个概率分布的距离,KL 值越大,表明主题与主题间的距离越远,主题质量越高。平均 KL 距离 avg_KL 的定义如公式(11)所示:

$$avg_KL=\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}KL(\varphi_i|\varphi_j)}{K^2} \qquad \text{式(11)}$$

其中 KL 距离 $KL(\varphi_i \parallel \varphi_j)=\sum_{v=1}^{V}\varphi_{iv}\log\frac{\varphi_{iv}}{\varphi_{jv}}$。由于 KL 距离是不对称的,但是 φ_i 和 φ_j 相似性度量是对称的,故将公式进行调整,采用对称的 Jensen-Shannon 距离度量 2 个主题词分布的距离,替代公式(11)中的 KL,具体计算公式如式(12)所示:

$$JS(\varphi_i,\varphi_j)=\frac{KL(\varphi_i,\varphi_j)+KL(\varphi_j,\varphi_i)}{2} \qquad \text{式(12)}$$

4.3 实验步骤与参数设置

首先进行术语识别。为此,使用中国科学院计算研究所的 ICTCLAS 分词系统(http://ictclas.nlpir.org/)对目标专利文本集和辅助专利文本集进行分词。该分词系统具备中文分词、词性标注等功能,是当前较好的中文分词工具,被广泛使用[20]。根据分词信息,计算目标集中单词的类别熵,通过人工判定,选取前 500 个最高的类别熵词作为通用词,对目标集进行粗切分,选取候选术语,然后使用 C-value 值排序候选术语选取最大的若干候选术语作为术语。

表 2 为类别熵值最高的前 10 个词。由表 2 可见,类别熵值最高的这些词通常在各专利文献类别中均会出现,与具体专利主题分析中的专利术语无关,包含语义信息较少,可以作为通用词。

表 2 类别熵值最大的前 20 个词

序号	词	类别熵
1	公开	3.00
2	发明	2.99
3	涉及	2.99
4	技术	2.99
5	需要	2.99
6	能够	2.98
7	提供	2.98
8	快速	2.98
9	进行	2.98
10	特征	2.98

对识别出的术语,采用最大长度匹配方法,对目标专利文本被切分成碎片的词组进行分词优化后,形成术语包(bag-of-terms),使用主题模型建模。在主题建模过程中,根据经验设置 α = 50/K、β = 0.01,Gibbs 采样迭代次数参数为 2 000,保存迭代参数为 1 000。主题数 K 的选取通过计算基本专利主题模型(第 3.2 部分)的困惑度选取最优值,采用五折交叉验证。根据计算,实验设定稀土钢数据集的主题数 $K=15$、电解锰数

据集的主题数 $K=20$。

4.4 术语识别方法评估

为了验证本文提出的术语识别方法的有效性,实验使用以下两种方法进行比较:

(1)Rule-C-value:传统的基于术语构词规则匹配的方法选名词构成的短语,将词性匹配的名词术语抽取出来[7],然后使用C-value排序候选术语。

(2)EC-C-value:使用本文3.1部分提出的类别熵方法选取候选术语,然后根据候选术语C-value值降序排序,抽取候选术语。

实验结果如图2、图3所示。由图2、图3可见,在两个数据集中,EC-C-value方法显著好于Rule-C-value方法,表明通过类别熵选取候选术语与通过C-value值排序候选术语比规则匹配的方法更加有效。

表3列出了两种方法前10个候选术语,其中粗体表示正确抽取的候选术语。由表3可见,基于术语构词规则的候选术语选取方法不能有效过滤通用词,这些词高频出现,且不在停用词表中,产生许多错误候选术语,从而导致最后术语抽取准确率很低;而基于通用词的方法则能够将"制备""方法""发明""生产"等作为通用词,将这些高频词排除在候选术语之外,从而提高了术语抽取的准确率。

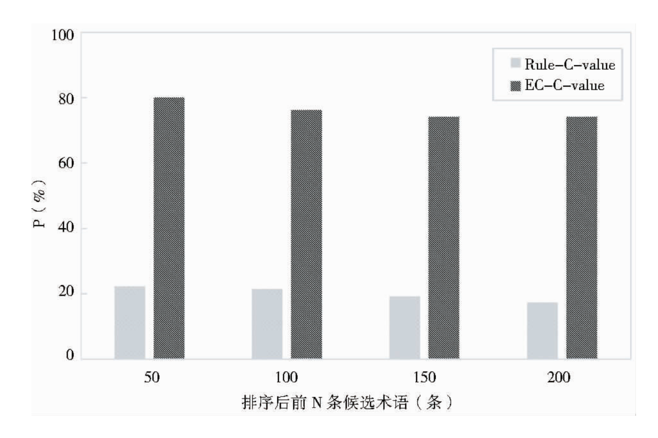


图2 稀土钢数据集术语识别方法比较

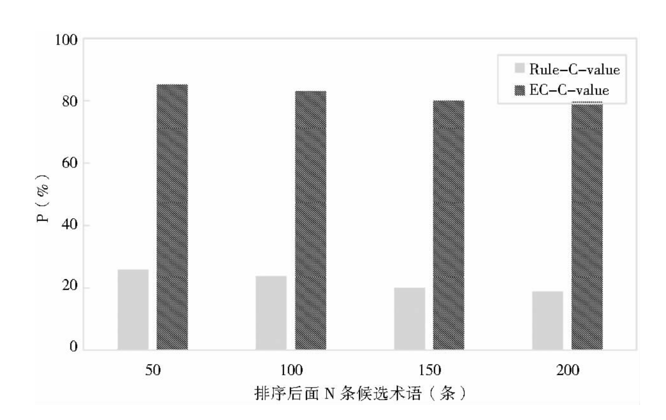


图3 电解锰集术语识别方法比较

表3 两种方法前10个候选术语

序号	稀土钢		电解锰	
	Rule-C-value	EC-C-value	Rule-C-value	EC-C-value
1	制备 方法	双向 不锈钢	制备 方法	电解 金属锰
2	发明 公开	稀土 永磁	阴极 板	药芯 焊丝
3	重量 百分比	稀土 合金	电解 金属锰	金属 铬
4	质量 百分比	合金 元素	电解锰 渣	金属锰 粉
4	含 稀土	奥氏体 不锈钢	药芯 焊丝	药 芯
5	制造 方法	结 硬质合金	重量 份	氧化 锰矿
6	生产 方法	稀土 氧化物	电解锰 阴极	锰 酸 锂
7	钢 结	钢 结 硬质合金	重量 百分比	碳酸 锰
8	技术领域	废旧 磁钢	电解锰 阴极 板	不锈钢 焊条
9	稀土 永磁	药芯 焊丝	硫酸锰 溶液	熔敷 金属
10	稀土 合金	铁素体 不锈钢	阳极 液	钝化 液

4.5 主题模型评估

为了验证本文提出的基于GPU的主题模型的有效性,实验使用第3.1部分方法选取的术语,对比以下两种主题模型:

(1)EC-PhraseLDA:该模型认为多词出现的连续词语并非碰巧,而这些连续的单词之间必然有一定的联系。使用势函数表示术语中词语之间的相互影响关系,术语的单词共享相同的潜在主题[17]。

(2)EC-GPULDA:使用本文第3.2.2部分提出的基于GPU的主题发现模型对术语包建模。

实验结果如图4所示。由图4可见,EC-GPULDA结果好于EC-PhraseLDA。EC-PhraseLDA使用术语替代了分词的碎片单词组,但是同时也造成术语共现降低,从而影响了主题建模效果;本文提出的EC-GPUL-

DA 引入 GPU 模型,提高相似术语的采样概率,从而增加了主题模型的距离。

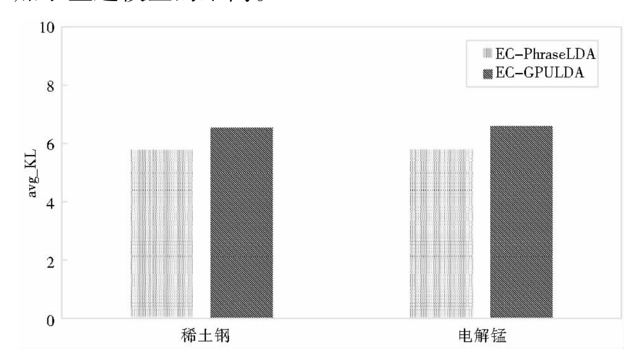


图 4　两种主题模型方法比较

表 4 列举了两种主题建模的前 10 个主题词和术语,其中多词术语使用粗体表示。由表 3 可以看出,使用 EC-PhraseLDA 时,由于多词术语稀疏性,使得前 10 个主题词中不包含多词术语;而 EC-GPULDA 方法能够根据相似度将相似的术语推论主题,从而使得一些较稀疏的多词术语能够进入前 10 单词,从而解决了术语稀疏性问题。

表 4　主题表示实例

序号	稀土钢		电解锰	
	EC-PhraseLDA	EC-GPULDA	EC-PhraseLDA	EC-GPULDA
1	焊接	焊接	电解锰	电解锰
2	焊条	钢筋	添加剂	电解
3	钢筋	药皮	电解	添加剂
4	渣	低合金钢	离子交换	氯化 锰
5	药皮	熔覆 金属	吸附	废液
6	钢芯	焊缝	硒	盐酸
7	焊缝	管线 钢	液体	聚 丙烯酰胺
8	焊芯	钢芯	废液	活性剂
9	钛铁	稀土 硅铁	导轨	洗涤器
10	氢	焊条	机械手	导轨

4.6　与传统专利主题比较

最后,实验将本文提出的方法与常用的专利主题分析方法进行比较。预比较的方法如下:

(1)LDA:去除专利文本中的停用词,对分词等预处理后的专利文本按照词为单位进行主题建模[5,9]。

(2)RuleLDA:根据词法规则,选取名词专利术语,术语中的单词属于相同的主题,进行主题模型建模[7]。

(3)EC-GPULDA:使用本文提出的方法,使用 EC 选取通用词,以识别专利术语,然后使用基于 GPU 的主题模型建模。

图 5 为实验结果。由图 5 可见,LDA 方法结果最差,RuleLDA 结果次之,本文提出的 EC-GPULDA 模型取得了最好的建模效果。

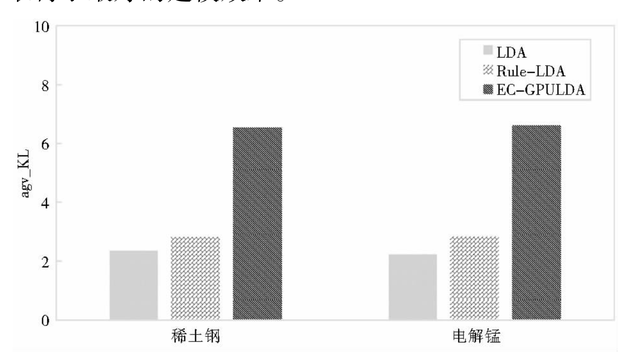


图 5　与传统专利主题模型比较

表 5 列举了 3 种主题模型中一个主题的前 10 个主题词和术语,其中多词术语使用粗体表示。由表 3 不难看出,基于词表示的主题模型包含一些通用词,如“发明”“所述”等,从而影响主题模型建模质量;而基于词性规则的主题模型在术语识别上准确率较低,如“生产 方法”并非真正的术语;以术语为主题表示的基本单元包含正确的多词术语,并通过 GPU 解决多词术语的稀疏性问题,使得模型更易于理解。

表 5　主题表示实例

序号	稀土钢			电解锰		
	LDA	RuleLDA	ECG-PULDA	LDA	RuleLDA	EC-GPULDA
1	发明	金属	焊接	所述	电解锰	电解锰
2	焊接	稀土	钢筋	连接	渣	电解
3	用于	组成	药皮	设有	环境	添加剂
4	焊条	力学性能	低合金钢	装置	方法	氯化 锰
5	金属	含量	熔覆 金属	电解锰	电解锰 渣	废液
6	钢	焊丝	焊接	电解槽	步骤	盐酸
7	焊丝	成型	管线 钢	设置	废水	聚 丙烯酰胺
8	重量	产品	钢芯	一种	锰渣	活性剂
9	药芯	特点	稀土 硅铁	包括	资源	洗涤器
10	包括	生产 方法	焊缝	管道	问题	导轨

俞琰,赵乃瑄.融入术语知识的专利主题发现方法[J].图书情报工作,2018,62(21):118-126.

5 结语

针对专利主题发现中,专利术语被切分为碎片,导致主题难以解释的问题,本文提出融入术语知识的专利主题模型。该模型首先根据专利文献的特点,引入类别熵,有效地识别出中文专利文献中的术语;然后利用泛化波利亚瓮模型加大语义相似术语分配到同一主题的概率,以缓解术语作为基本主题模型分析单位所带来的数据稀疏性问题。模型将术语作为主题模型分析的基本单位,同传统的基于分词的主题模型表示方法相比,术语包含更加丰富的语义信息、具有更强的可读性。实验结果表明,基于术语的表示方法明显提高了主题的可读性,从共现和语义相关两个角度有效降低了术语稀疏性的影响。模型不需要领域知识和繁杂的语言规则,属于数据驱动的算法,适合于大规模中文专利文献进行主题分析。

虽然本文融入术语知识的专利主题模型较传统专利主题模型取得了更好的效果,但是专利术语识别中低频术语识别的准确率不高,在用 C-value 排序后,低频候选术语 C-value 值较小而被移除。在将来的工作中,将进一步深入研究如何提高低频术语识别的准确率。

参考文献:

[1] TANG J, WANG B, YANG Y, et al. PatentMiner: topic-driven patent analysis and mining[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012:1366-1374.

[2] WANG B, LIU S, DING K, et al. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1):685-704.

[3] CHEN H, ZHANG G, LU J, et al. A fuzzy approach for measuring development of topics in patents using Latent Dirichlet Allocation [C]// IEEE international conference on fuzzy systems. Piscataway, NJ: IEEE, 2015:1116-1116.

[4] KIM M, PARK Y, YOON J. Generating patent development maps for technology monitoring using semantic patent-topic analysis[J]. Computers & industrial engineering, 2016, 98(1):289-299.

[5] SUOMINEN A, TOIVANEN H, SEPPANEN M. Firms' knowledge profiles: mapping patent data with unsupervised learning[J]. Technological forecasting & social change, 2016, 115(1):1-12.

[6] 范宇, 符红光, 文奕. 基于 LDA 模型的专利信息聚类技术[J]. 计算机应用, 2013, 33(S1):87-89.

[7] 王博, 刘盛博, 丁堃, 等. 基于 LDA 主题模型的专利内容分析方法[J]. 科研管理, 2015, 36(3):111-117.

[8] 吴菲菲, 张亚茹, 黄鲁成, 等. 基于 AToT 模型的技术主题多维动态演化分析——以石墨烯技术为例[J]. 图书情报工作, 2017,1(5):95-102.

[9] 廖列法, 勒孚刚. 基于 LDA 模型和分类号的专利技术演化研究[J]. 现代情报, 2017, 37(5):13-18.

[10] 陈亮, 张静, 张海超, 等. 层次主题模型在技术演化分析上的应用研究[J]. 图书情报工作, 2017,1(5):103-108.

[11] WALLACH H M. Topic modeling: beyond bag-of-words[C]// International conference on machine learning. New York: ACM, 2006:977-984.

[12] WANG X, MCCALLUM A, WEI X. Topical N-grams: phrase and topic discovery, with an application to information retrieval[C]// IEEE international conference on data mining. Piscataway, NJ: IEEE, 2007:697-702.

[13] LINDSEY R V, Headden III W P, STIPICEVIC M J. A phrase-discovering topic model using hierarchical Pitman-Yor processes [C]// Joint conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg, PA: ACL,2012:214-222.

[14] DANILEVSKY M, WANG C, DESAI N, et al. Automatic construction and ranking of topical keyphrases on collections of short documents[C]// Proceedings of the 2014 SIAM international conference on data mining. Philadelphia, PA: SIAM, 2014: 398-406.

[15] El-KISHKY A, SONG Y, VOSS C R, et al. Scalable topical phrase mining from text corpora[J]. Proceedings of the VLDB endowment, 2014, 8(3):305-316.

[16] 张琴, 张智雄. 基于 PhraseLDA 模型的主题短语挖掘方法研究[J]. 图书情报工作, 2017,61 (8):120-125.

[17] HEINRICH G. A generic approach to topic model[M]//Machine learning knowledge discovery in databases. Berlin: Springer, 2009:517-532.

[18] ZIPF G K. Selected studies of the principle of relative frequency in language[J]. Language, 1933, 9(1):89-92.

[19] 韩红旗, 朱东华, 汪雪锋. 专利技术术语的抽取方法[J]. 情报学报, 2011, 30(12):1280-1285.

[20] 徐川, 施水才, 房祥, 等. 中文专利文献术语抽取[J]. 计算机工程与设计, 2013, 34(6):2175-2179.

[21] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms:. the C-value/NC-value, method[J]. International journal on digital libraries, 2000, 3(2):115-130.

[22] SPASIC I, GREENWOOD M, PREECE A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 4(1):27-42.

[23] MAYNARD D, ANANIADOU S. Identifying terms by their family and friends [C]// Conference on computational linguistics.

Stroudsburg, PA:ACL, 2000:530-536.
[24] 李超, 王会珍, 朱慕华,等. 基于领域类别信息 C-value 的多词串自动抽取[J]. 中文信息学报, 2010, 24(1):94-99.
[25] 刘里, 刘小明. 基于分隔符和上下文术语的领域现象术语抽取[J]. 华南理工大学学报(自然科学版), 2011, 39(7):146-149.
[26] 胡阿沛, 张静, 刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术, 2013, 29(2):24-29.
[27] 张杰, 张海超, 翟东升. 面向中文专利权利要求书的分词方法研究[J]. 现代图书情报技术, 2014, 30(9):91-98.
[28] MAHMOUD H. Polya urn models[M]. New York: Champman & Hall/CRC, 2009.
[29] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1):993-1022.
[30] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]// Proceedings of the national academy of Science, 2004, 1(1):5228-5235.
[31] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]// Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA:ACL, 2011:262-272.
[32] CHEN Z, MUKHERJEE A, LIU B, et al. Leveraging multi-domain prior knowledge in topic models[C]// International joint conference on artificial intelligence. Menlo Park, CA: AAAI, 2013: 2071-2077.
[33] CHEN Z, MUKHERJEE A, LIU B, et al. Discovering coherent topics using general knowledge[C]// ACM international conference on information & knowledge management. New York: ACM, 2013:209-218.
[34] CHEN Z, LIU B. Mining topics in documents: standing on the shoulders of big data[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM, 2014: 1116-1125.
[35] 孙锐, 郭晟, 姬东鸿. 融入事件知识的主题表示方法[J]. 计算机学报, 2017, 40(4):791-804.

作者贡献说明：

俞琰:提出研究思路,设计研究方案,进行实验,负责论文撰写;

赵乃瑄:负责分析数据,修改论文。

Patent Topic Discovery Method Integrated with Term Knowledge

Yu Yan[1,2] Zhao Naixuan[1]

[1] Information Service Department, Nanjing Tech University, Nanjing 210009

[2] Computer Science Department, Southeast University Chengxian College, Nanjing 211816

Abstract: [**Purpose/significance**] Aiming at the problem of analysis patent topic in terms of word which causes topics are difficult to explain in the patent topic analysis, this paper proposes a patent topic discovery model integrated with term knowledge. [**Method/process**] The proposed model firstly introduces the class entropy and effectively recognizes the terms in the patent literature. Then, the Generalized Pólya Urn model is used to increase the probability of the semantic similarity terms assigned to the same topic, in order to alleviate the data sparsity problem brought by the term as the basic topic model analysis unit. [**Result/conclusion**] The experimental results show that the proposed model contains the term information to improve the quality of the topic generation, making the topic representation more readable and topic discriminative.

Keywords: patent analysis topic discovery term